Modèles et outils pour la publication de métadonnées d'archives géographiques et de leurs données dérivées

Melvin Hersent, Abadie Nathalie, Duménieu Bertrand, Perret Julien

Modèles et outils pour la publication de métadonnées d'archives géographiques et de leurs données dérivées

Hersent Melvin[1], **Abadie Nathalie**[1], **Duménieu Bertrand**[2], **Perret Julien**[1]

[1] LASTIG, Univ. Gustave Eiffel, IGN-ENSG, F-94160 Saint-Mandé, France
{melvin.hersent, nathalie-f.abadie, julien.perret}@ign.fr
[2] CRH-EHESS, Paris, France
bertrand.dumenieu@ehess.fr

Résumé

L'interopérabilité des données dans un projet pluridisciplinaire est primordiale. Prenant l'exemple d'un projet de recherche en histoire spatiale, nous comparerons dans un premier temps les standards et vocabulaires à notre disposition pour décrire des données géographiques et des documents d'archives. Nous proposons ensuite un alignement entre les standards retenus : l'ISO 19115 et RiC-O. Enfin, nous proposons une architecture de micro-services pour la saisie, le stockage, la publication sur le Web et l'interrogation unifiée des métadonnées de nos sources.

1 Introduction

La réutilisabilité des données produites par la recherche est un sujet important, qui suppose que les données soient ouvertes et documentées par des métadonnées appuyées sur des vocabulaires partagés et eux-mêmes ouverts. Les principes FAIR (Wilkinson et al., 2016) indiquent que pour répondre de manière efficace à ces besoins les données produites doivent être : **F**acile à trouver, **A**ccessibles, **I**nteropérables et **R**éutilisables [1].

Dans le champ des humanités numériques, les projets de recherche en histoire spatiale présentent la particularité d'exploiter des sources géographiques, telles que des archives textuelles à caractère géographique (ex. des annuaires d'adresses) ou des cartes et plans topographiques. Ces projets produisent des données géohistoriques comme des cartes et plans géoréférencés, et des données géographiques vectorielles [2]. De l'application sur ces données de processus d'extraction d'information, de transformation ou de liage dérivent, parfois récursivement, de nouvelles données qui sont elles-mêmes à décrire.

Les caractéristiques géographiques des sources comme de leurs données dérivées sont primordiales et il faut permettre aux chercheurs de décrire celles-ci par un système de coordonnées, une emprise spatiale, une résolution spatiale, etc., tout en respectant les principes FAIR. L'objectif est donc de faciliter la documentation de ces deux types de données et de faciliter leur découverte et leur accès.

Dans cet article nous passons en revue les modèles de métadonnées adaptés aux documents d'archives à caractère géographique et aux données géographiques numériques produites à partir de ces sources. Nous identifions ensuite les correspondances, les divergences et les complémentarités des deux modèles retenus et nous proposons un ensemble de mappings et d'extensions permettant la conversion des métadonnées de l'un vers l'autre afin d'assurer leur interopérabilité. Enfin, nous proposons une architecture de micro-services pour la saisie, le stockage, la publication sur le Web et l'interrogation unifiée des métadonnées de nos sources.

2 Standards et vocabulaires pour les données géographiques et les documents d'archives

Dans le domaine de la géomatique, la solution recommandée pour décrire et diffuser des données géographiques est l'utilisation d'une Infrastructure de Données Géographiques (IDG). Une IDG désigne une collection pertinente de technologies, de politiques et de dispositions institutionnelles pour permettre la disponibilité et faciliter l'accès à des données spatiales (Nebert, 2004).

Suivant la Directive Européenne INSPIRE [3], la

1. https://www.go-fair.org/fair-principles/

2. La forme et la localisation des entités géographique est encodée par une géométrie discrète faite de primitives géométriques, points, polylignes et polygones, dont les coordonnées sont exprimées dans un système de référence connu.

3. https://inspire.ec.europa.eu/

norme ISO 19115 [4] définie par l'ISO/TC 211 [5] doit être privilégiée pour décrire des jeux de données géographiques dans un catalogue d'IDG. Cette norme est particulièrement adaptée pour décrire des données géographiques institutionnelles et fournit les éléments nécessaires à la documentation de leur système de coordonnées de référence, leur emprise et leur résolution spatiales, le type de géométrie utilisé et le format de données adopté, tout en étant extensible aux besoins spécifiques d'autres domaines (Organisation internationale de normalisation, 2014). En revanche, les catalogues des IDG ne respectent pas les principes FAIR par nature (Tandy et al., 2017). Il manque pour chaque jeu de données : une URI unique et pérenne, d'être indexable par les moteurs de recherche, de lier les ressources entre elles et d'exposer les données grâce à une API.

En France, les recommandations institutionnelles dans le domaine culturel sont de s'appuyer sur les bonnes pratiques du Web de données [6] pour publier des métadonnées sur le Web (Ministère de la Culture et de la Communication, 2014). Pour les données d'archives, trois modèles et les ontologies associées ont été envisagés : *Dublin-Core* [7], *Cidoc Conceptual Reference Model* [8] (Cidoc-CRM) et *Records in Context - Conceptual Model* (RiC-CM) [9]. Dublin-Core fournit trop peu d'éléments de description des métadonnées pour nos besoins tandis que Cidoc-CRM est plus adapté à des métadonnées de collections muséales. De plus, les deux ne couvrent pas la notion d'algorithme utilisée sur les jeux de données, présente en ISO 19115 dans la catégorie *Lineage* et en RiC-CM avec le concept de *Mechanism*. RiC-CM correspond parfaitement à la nature de nos sources : nous devons décrire des documents d'archives dont plusieurs exemplaires peuvent coexister, conservés par différentes institutions patrimoniales et culturelles. C'est une information importante pour retracer la généalogie des données numériques produites à partir d'archives. Ainsi, le modèle RiC-CM permet de décrire un atlas ancien comme une instance de la classe *RecordSet* et chacune de ses feuilles comme autant d'instance de la même classe. Un exemplaire de cet atlas, conservé aux Archives Nationales par exemple, est alors une instance d'*Instantiation* de notre *RecordSet*, et chacune de ses pages, une instance d'*Instantiation* de la page de type *RecordSet* correspondante. Toute page scannée est alors une instance d'*Instantiation*, reliée à la précédente par la propriété *hasDerivedInstantation* (voir fig. 1). Cette distinction, entre un jeu de données et ses différentes distributions, n'est pas faite dans l'ISO 19115. Le modèle RiC-CM a été traduit en OWL(Group, 2012), le langage standard de représentation d'ontologies et l'ontologie RiC-O [10] correspondante possède déjà une extension pour décrire les caractéristiques d'archives géographiques [11].

3 Alignement de l'ISO 19115 et RiC-O

Nous avons effectué une comparaison entre les vocabulaires utilisés par différents catalogues de métadonnées : Zenodo [12], Nakala [13], CKAN [14], Dataverse [15], Didomena [16] et enfin Geonetwork [17], le seul IDG natif de cette liste. CKAN peut être utilisé comme IDG avec le vocabulaire GeoDCAT-AP [18], mais cela nécessite l'ajout d'extensions et complexifie le déploiement.

Tous utilisent des modèles de métadonnées prédéfinis et aucun ne permet de saisir et publier des métadonnées conformes à l'ontologie RiC-O. Geonetwork possède nativement des outils de saisie et de requêtes permettant de manipuler des informations géographiques et utilise la norme ISO 19115 pour décrire les métadonnées. Nous proposons donc de l'adopter pour disposer d'une interface de saisie et de manipulation des métadonnées conviviale. Pour nous conformer aux principes FAIR, nous ajoutons à notre infrastructure de publication de métadonnées l'outil Ontop [19] qui nous permet de mettre en place un graphe virtuel et de publier notre catalogue selon les bonnes pratiques du Web de données sans devoir déployer une seconde base de données orientée graphe.

4. https://www.iso.org/fr/standard/53798.html
5. https://www.iso.org/fr/committee/54904.html
6. https://www.w3.org/TR/sdw-bp/
7. https://www.bnf.fr/fr/dublin-core
8. https://www.cidoc-crm.org/
9. https://www.ica.org/fr/records-in-contexts-modele-conceptuel
10. https://www.ica.org/standards/RiC/ontology
11. http://data.alegoria-project.fr/def/geotheque#
12. https://zenodo.org/
13. https://nakala.fr/
14. https://ckan.org/
15. https://dataverse.org/
16. https://didomena.ehess.fr/
17. https://geonetwork-opensource.org/
18. https://inspire.ec.europa.eu/good-practice/geodcat-ap
19. https://ontop-vkg.org/guide

L'outil de graphe virtuel permet de convertir à la volée les métadonnées ISO 19115 stockées dans la base de données du catalogue Geonetwork, en RDF (Cyganiak et al., 2014), le modèle recommandé pour publier des données sur le Web, conformément à l'ontologie RiC-O. Cette étape suppose de lui fournir un ensemble de *mappings*, c'est-à-dire de règles de conversion permettant de passer d'un modèle à l'autre.

Pour faciliter l'écriture des *mappings* nous avons adapté le modèle proposé par Geonetwork. En ISO 19115, l'exemple de l'Atlas de Verniquet se résume au diagramme présenté en figure 1 : l'atlas est une instance de *Dataset* et ses feuilles des instances de *Resource*. Ceci permet de spécifier le traitement effectué sur chaque feuille, mais pas de distinguer et décrire différents exemplaires de l'atlas.

Pour réconcilier les deux modèles nous adaptons l'ISO 19115 de la manière suivante (figure 2) : nous créons une instance *DS_Aggregate* pour regrouper les collections de ressources (des instances de *DS_Dataset*) portant sur le même sujet, qui sont elles-mêmes composées d'instances de *DS_Resources*. Tout cela correspond à des instances du concept de *RecordSet* en RiC-CM et nous ajoutons un *keyword* (*RecordSet*) dans chaque jeu de métadonnées pour marquer cela [20]. Pour obtenir l'équivalent des *Instantiations* en RiC-CM, nous créons de nouveaux *DS_Datasets* et *DS_Resources* ayant pour source ceux possédant le mot clé *RecordSet*. Nous ajoutons également un *keyword* (*Instantiation*) pour les différencier. Les caractéristiques géographiques des données sont représentées en RDF à l'aide des ontologies de systèmes de référence de coordonnées [21] et de spécification de données géographiques [22].

4 Une infrastructure de publication de métadonnées d'archives géographiques et de leurs données dérivées

Pour notre infrastructure, représentée en figure 3, nous profitons d'une solution clé en main fourni par Geonetwork contenant le catalogue, une base de donnée spatiale (PostgreSQL avec l'extension PostGIS [23]) et une instance d'Elasticsearch accompagnée de Kibana [24] pour l'indexation des données. Cette solution utilise Docker [25] et nous assure une facilité de déploiement et la portabilité des applications et des données sauvegardées.

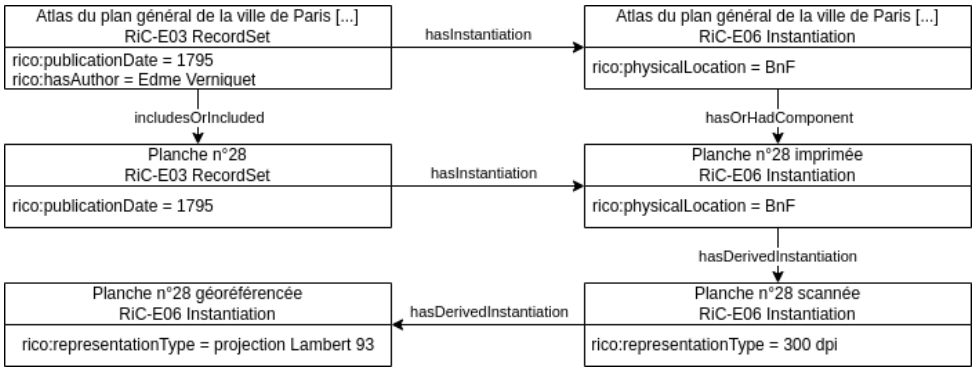
Nous connectons à notre base de donnée une instance de pgAdmin4 [26], un *dashboard* pour faciliter la manipulation des données et Ontop pour gérer les *mappings*, créer notre graphe virtuel, disposer d'un *endpoint* SPARQL et, grâce à une instance de Lodview [27], d'une interface permettant de visualiser les métadonnées RDF de manière conviviale. Nous pouvons voir une comparaison entre les interfaces de Geonetwork et de LodView sur la figure 4.

Les images de nos ressources sont exposées par le serveur Cantaloupe [28] à travers une API IIIF (Snydman et al., 2015) [29], standard de diffusion d'images largement utilisé par les bibliothèques numériques. Cela permet de déployer une application de visualisation d'atlas historiques [30] développée par le consortium Allmaps [31] sur un fond de carte moderne. Cette solution permet un géoréférencement des images au format IIIF à la volée à partir d'un fichier contenant des points de contrôle (format *georeference annotation* proposé au consortium IIIF pour devenir une extension officielle IIIF). Enfin, pour gérer les urls nous déployons une instance de Traefik [32], conçu spécialement pour fonctionner avec des services Docker.

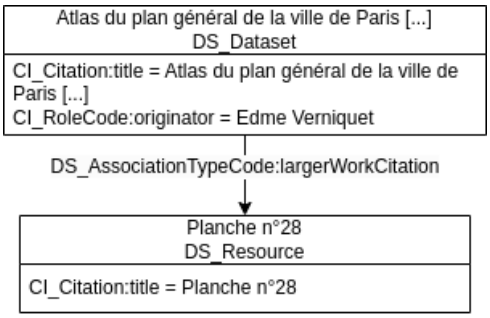
5 Conclusion

Nous avons comparé différents modèles et outils pour proposer une solution pour la découverte et l'accès à des sources géographiques historiques et leur données numériques dérivées produites lors de projets de recherche en histoire spatiale, ainsi que pour favoriser leur réutilisation. Notre approche combine un catalogue standard d'IDG, avec un graphe virtuel publiant les métadonnées produites sur le Web conformément au vocabulaire RiC-O. Ceci présente l'avantage de faciliter la description

20. Nous avons choisi de créer uniquement des instances de RecordSet, et pas Record, pour conserver la possibilité de représenter des archives cartographiques fragmentaires.
21. http://data.ign.fr/def/ignf#
22. http://data.ign.fr/def/xysemantics
23. https://postgis.net/
24. https://www.elastic.co/fr/elasticsearch/
25. https://github.com/geonetwork/docker-geonetwork
26. https://www.pgadmin.org/
27. https://lodview.it/
28. https://cantaloupe-project.github.io/
29. https://iiif.io/
30. https://github.com/soduco/webgl2-preview basée sur https://github.com/allmaps/webgl2-preview
31. https://allmaps.org/
32. https://doc.traefik.io/traefik/



(a) Représentation RiC-O



(b) Représentation ISO 19115

FIGURE 1 – Un exemple de plan ancien décrit avec le modèle RiC-CM (a) et avec la norme ISO 19115 (b) : l'Atlas général de la ville de Paris d'Edme Verniquet.

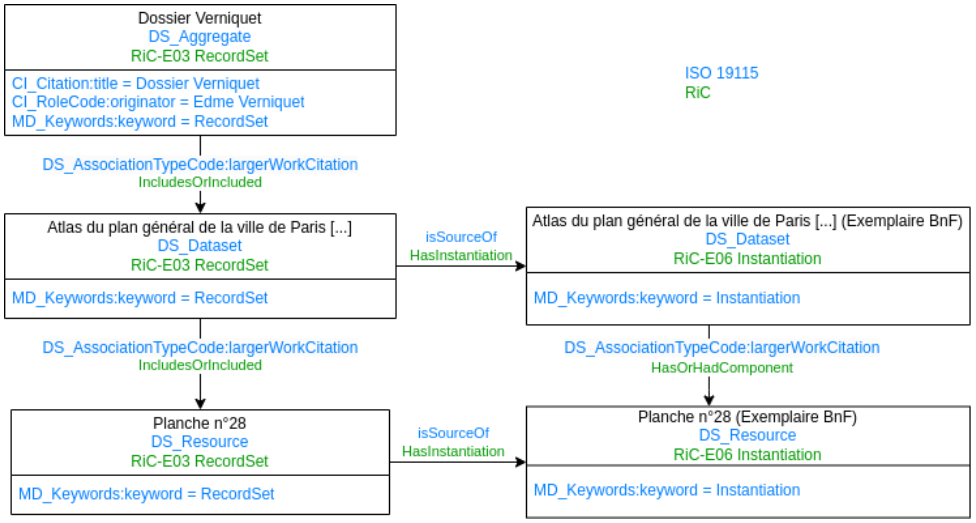


FIGURE 2 – Proposition pour rapprocher l'ISO 19115 et le modèle RiC-CM

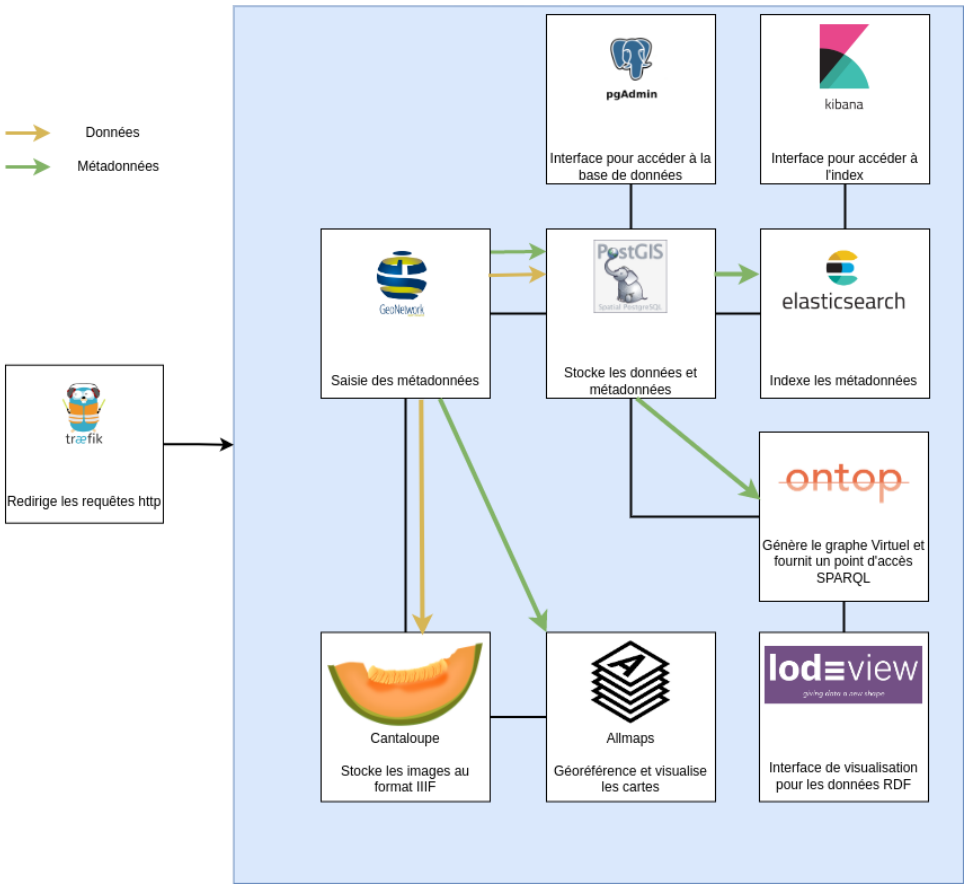


FIGURE 3 – Schéma de l'architecture déployée. Les relations fonctionnelles sont représentées par les liens en noir, les flux de données et de métadonnées par les flèches jaunes et vertes.

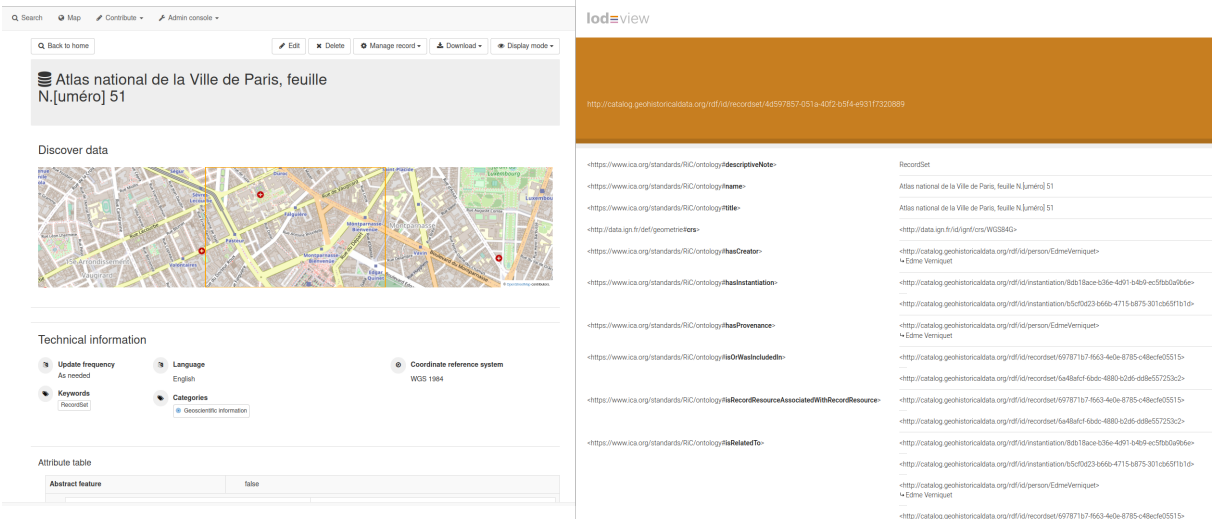


FIGURE 4 – Comparaison entre l'affichage des métadonnées sur Geonetwork (à gauche) et celui de leur conversion en RDF, produit par LodView (à droite)

des caractéristiques géographiques des données et de permettre leur découverte conformément aux principes FAIR. L'infrastructure mise en place a été testée sur quelques exemples de cartes anciennes et les données numériques dérivées (scans et données vecteurs). À terme, nous souhaitons étendre les métadonnées produites pour décrire les relations de filiation entre cartes anciennes et faciliter la découverte de corpus d'archives cartographiques géométriquement cohérentes.

Remerciements

Ce travail a été soutenu financièrement par l'Agence Nationale de la Recherche dans le cadre du projet SODUCO (ANR-18-CE38-0013).

Bibliographie

Richard Cyganiak, David Wood, et Markus Lanthaler. 2014. Rdf 1.1 concepts and abstract syntax. Rapport technique, W3C.

W3C OWL Working Group. 2012. Owl 2 web ontology language - document overview (second edition). Rapport technique, W3C.

Ministère de la Culture et de la Communication. 2014. Feuille de route strategique : Métadonnées culturelles et transition web 3.0. Rapport technique 2014-01, Ministère de la Culture et de la Communication, 182, rue Saint-Honoré, 75033 Paris Cedex 01.

Douglas D Nebert. 2004. Developping spatial data infrastructures: The sdi cookbook. Rapport technique, Technical Working Group Chair, GSDI.

Organisation internationale de normalisation. 2014. Iso 19115-1:2014 information géographique — métadonnées — partie 1: Principes de base. Rapport technique 2014-04, Organisation internationale de normalisation, Chemin de Blandonnet 8 CP 401 - 1214 Vernier, Geneva, Switzerland.

Stuart Snydman, Robert Sanderson, et Tom Cramer. 2015. The international image interoperability framework (iiif) : A community & technology approach for web-based images. In *Archiving conference*, 1, pages 16–21. Society for Imaging Science and Technology.

Jeremy Tandy, Linda van den Brink, et Payam Barnaghi. 2017. Spatial data on the web best practices. Rapport technique OGC 15-107, Open Geospatial Consortium & W3C.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1) :1–9.